DOCUMENT RESUME

ED 064 363 TM 001 531

AUTHOR Huberty, Carl J.; Blommers, Paul J.
TITLE An Empirical Comparison of Three Indices of Variable Contribution in Multiple Group Discriminant Analysis.
PUB DATE 72
NOTE 18p.; Paper presented at the 1972 AERA conference

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; Correlation; *Discriminant Analysis; *Predictive Validity
IDENTIFIERS *Fisher Discriminant Function

ABSTRACT

An empirical comparison of three proposed indices of predictor variable potency is presented. These indices are: (1) the scaled weights of the first Fisher-type discriminant function, (2) the total group estimates of the correlations between each predictor variable and the first Fisher-type function, and (3) the within-groups estimates of the correlations between each predictor variable and the first Fisher-type function. It was found that given a single run of an experiment none of the indices were sufficiently reliable to be of great practical value in identifying potent variables except when the total sample size was very large. (Author)

ED 064363

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

AN EMPIRICAL COMPARISON OF THREE INDICES OF VARIABLE CONTRIBUTION IN MULTIPLE GROUP DISCRIMINANT ANALYSIS

Carl J Huberty
University of Georgia

Paul J. Blommers
University of Iowa

TM 001 531

AERA, 1972

Abstract

This paper is concerned with an empirical comparison of three proposed indices of predictor variable potency: (1) the scaled weights of the first Fisher-type discriminant function, (2) the total group estimates of the correlations between each predictor variable and the first Fisher-type function, and (3) the within-groups estimates of the correlations between each predictor variable and the first Fisher-type function. It was found that given a single run of an experiment none of the indices were sufficiently reliable to be of great practical value in identifying potent variables except when the total sample size was very large.

AN EMPIRICAL COMPARISON OF THREE INDICES OF
VARIABLE CONTRIBUTION IN MULTIPLE GROUP DISCRIMINANT ANALYSIS

R. A. Fisher's original discriminatory technique is well known and involves the reduction of a set of p measures to a single (composite) measure, a "linear discriminant function" (LDF),

$$Y = w_1 X_1 + w_2 X_2 + \ldots + w_p X_p$$

The w_i ($i = 1, \ldots, p$) are determined so as to maximize the ratio of the differences between the sample means of the Y-values to the standard error of this difference as estimated from the within-groups mean square. In analysis of variance terminology this is equivalent to finding the w_i such that, for the Y-values, the ratio

$$\frac{MS_{bet}}{MS_w}$$

is maximized. The popularity of the extension of Fisher's two-group analysis to one involving k groups is principally due to the efforts of statisticians at Harvard University.

As described by Cooley and Lohnes (1962 Ch. 6) multiple group discriminant analysis strategy begins with a principal components

analysis. This analysis is made, not of the predictor variable intercorrelation matrix, but of the matrix product $E^{-1}H$, where E and H are the pooled within-groups and the among-groups deviation score cross-products matrices, respectively. This "factoring" may be construed as a partitioning of the discriminatory power of the set of *p* predictor variables into uncorrelated components called discriminant functions. The vectors obtained from the eigenanalysis of $E^{-1}H$ define a (discriminant) space such that when points representing the groups are located within it, these points are separated from each other to a maximum degree.

The sample estimates of the coefficients of the discriminant functions are determined by the eigenvectors, $\underline{u}_i$, associated with the eigenvalues, λ_i, of the determinantal equation

$$|E^{-1}H - \lambda I| = 0.$$

The eigenequation which leads to the coefficients is

$$(E^{-1}H - \lambda_i I)\underline{u}_i = 0.$$

This latter equation is obtained as a result of maximizing the ratio of mean-square among to the mean-square within groups. The maximum number of discriminant functions necessary to represent group differences is the smaller of p and k - 1.

To find the dimension of the so-called "discriminant space" either the eigenvalues are subjected to a significance test (see Rao, 1952, pp. 372-373), or a subset of the non-zero eigenvalues that accounts for a large percent, say 80, of the total discriminating power of the predictor variables may be chosen. The obtained eigenvectors may be normalized, and the elements of these normalized eigenvectors are then the coefficients of the discriminant functions. These normalized vectors may be scaled by multiplying each element by the square root of the corresponding variance estimate obtained from the principal diagonal of matrix E.

In many situations involving multiple group discriminant analysis interest is centered on the relative contribution of each predictor variable to the discrimination involved. By analyzing the k group centroids in the reduced or discriminant space, it is possible to determine the role of each of the discriminant functions retained. That is, some insight into the question, "Between what groups or sets of groups does each function discriminate?" may be gained. It is informative to determine which predictor variables are contributing the most to such discriminations. The problem of determining indices of variable potency in terms of contribution to discrimination is one which has plagued researchers for some time.

One possible solution of this problem has been proposed by Cooley and Lohnes (1962, p. 118). They claim that the scaled coefficients (i.e., "beta" or "standardized" weights) mentioned above indicate the relative contribution of each variable in determining the discriminant score.

Cautions against the use of such weights are pointed out by Bock and Haggard (1966, pp. 118-119).

Another approach to the problem of assessing the predictor variable contribution to discrimination involves a consideration of estimates of the correlations between each of the predictors and each of the discriminant functions. The matrix of such correlations constitutes what is known in factor analytic terminology as a "structure" matrix. Two approaches have been used to compute these estimates. When the data collected are considered as representing a single population, the structure matrix is determined as follows. Let q be the dimension of the determined discriminant space. Then the (q x q) diagonal matrix of the variances of the scores on each of the discriminant functions is given by

$$G = U' D_T U ,$$

where

U = the (p x q) matrix of orthogonal normalized eigenvectors of $E^{-1}H$, and

D_T = the (p x p) total covariance matrix.

The discriminant weights are then rescaled by the following matrix multiplication:

$$B = UG^{-1/2} .$$

Here B is a (p x q) matrix. Let S denote a (p x p) diagonal matrix of the variances of the predictor variables (i.e., of the elements of the main diagonal of D_T). Then the (p x q) matrix, Z, of discriminant weights appropriate for use with standardized scores is

$$Z = S^{1/2}B.$$

The (p x q) structure matrix is then given by

[1] $$A = RZ,$$

where R is the (p x p) predictor variable intercorrelation matrix. Correlations computed this way are precisely the r-values that would result if the Pearson product-moment coefficients between the sample predictor scores and the sample discriminant scores were calculated (see Gulliksen (1950, p. 339). In this study only the correlations between the predictors and the "leading" discriminant function were considered.[1] Therefore, only the first column of A in [1] is of interest here.

If the underlying model is one of k populations with identical covariance matrices, then the maximum likelihood estimate of the true correlation vector is given by (Bargmann, 1970, p. 53 or Porebski, 1966, p. 225)

[2] $$\underline{r}^* = (\underline{v}\ E\ \underline{v}')^{-\frac{1}{2}}\ (\underline{v}\ E)\ D_{1/\sqrt{e_{ii}}},$$

where

$\underline{v}$ = (1 x p) vector of weights for the first discriminant function, and

$D_{1/\sqrt{e_{ii}}}$ = (p x p) diagonal matrix of the reciprocals of the positive square roots of the diagonal elements of E.

The purpose of this study was to investigate the stability, over repeated sampling, of three indices of predictor variable potency:

#1 The scaled discriminant function coefficients,

#2 The correlations from the first column of the structure matrix given by equation [1], and

#3 The correlations determined as in equation [2].

The computerized simulation procedure was developed for drawing random samples of size N from k p-variate normal populations having a known common covariance matrix. This sampling experiment was repeated to provide data for empirically checking the reliability (in the sense of consistency) and validity of the three indices studied. A more detailed description is provided elsewhere (Huberty, 1969).

Data Analysis

The criterion used to judge the stability of the three indices of predictor variable potency was the consistency of the observed rank of each variable, as determined by the absolute value of each index, over repeated replications of the experiment. A necessary but not sufficient essential of a valid index of variable potency is that it exhibits consistency over repeated sampling. That is, an index lacking such consistency provides no basis for inferential statements concerning

the respective worth of selected predictor variables. Only the rankings of the variables with respect to the first discriminant function were determined. These potency rankings were analyzed in two ways. First, the number of times each variable attained a given rank was determined. These number-of-times-per-rank counts were found for each of the N-values and each of the k-values studied. These counts were organized into 24 (3 indices x 4 values of N x 2 values of k) two-way tables, the rows corresponding to the 10 possible ranks and the columns corresponding to the 10 variables.

The second analysis involved a correlational approach. For each index the ranks of the variables, with respect to the first discriminant function, were found for each replication of the experiment. Ranks from 1 to 10 were assigned according to the numerical value of the index. Thus for each index a 100-by-10 two-way table was formed for each value of N and each value of k. The relationship among the 100 rankings was determined by computing the coefficient of concordance, W (Kendall, 1955, p. 95). This coefficient was computed for the first discriminant function, for each of the three indices, and each of the four values of N in both the three- and five-group situations. This resulted in the computation of 24 coefficients in all. The significance of each observed value of W, i.e., the hypothesis that there was no consistency in the rankings over the 100 replications, was tested using a chi-square statistic (Kendall, 1955, p. 98). When an observed value of W was found to be significant, i.e., when there was evidence of some agreement of the

potency of the discriminator variables over repeated sampling, an estimate of the true ranking was obtained by ranking the variables according to the sums of the ranks allotted over the 100 replications.[2]

Results

Three Group Case (k = 3). It was clear from tables[3] exhibiting the number of times each variable attained a given rank for each index that the stability of the indices over repeated sampling is not very marked. If an index is operating consistently over repeated sampling, then each column of such a table would contain only one value which is large in relation to the others; such a pattern was not observed.

The W-values and the observed chi-square values corresponding to them are given in Table 1. All of the values were significantly different from zero (at the .01 level). Some (when N = 90) of the observed W-values which, according to Kendall may be roughly interpreted as correlation (here in the sense of reliability) coefficients, are quite low. That these low values were significant is simply a result of the power of the test to detect differences between the hypothetical zero value of the population W-values, and their observed values which differences are of no practical consequence. It is clear that unless sample size is very large neither the scaled

---

Insert Table 1

---

weights nor either of the correlation estimates of variable potency are very consistent over repeated sampling.

The population weights for Variables 9 and 10 were arbitrarily fixed at zero (see Huberty, 1969). That is, these two variables would be expected to exhibit minimum potency insofar as their contribution to discrimination among groups is concerned. Hence, it was possible to effect, to some extent, an evaluation of the validity of the three indices under consideration.

While none of the indices provides a very reliable rank-order of variable potency for a single run of the experiment, the reliability of each index is nevertheless sufficient to provide a reliable (in the sense of consistent) estimate of variable potency when the ranks are averaged over 100 replications of the experiment. Table 2 gives the potency of each variable based on the average value of its rank as assigned by each index over 100 runs of the experiment. With one exception which occurred in the case of the smallest sample size ($N = 90$), Index #1 assigned potency ranks of least and next to least to Variables 9 and 10. Index #2, on the other hand, assigned potency ranks ranging from 4 to 6 to Variable 10 and the ranks assigned to this variable by Index #3 ranged from 4 to 7. Judged in the light of this criterion, Index #1 is clearly the most valid of the three.

As a check on the reliability of the average potency ranks over 100 replications of the experiment, Kendall's W was calculated for each index using the (average) ranks for the four sample sizes as

---

Insert Table 2

---

given in Table 2. The W-values for the three indices were .95, .92, and .91, respectively. When the sum of the (average) ranks over the four sample sizes is used as a basis for assigning an overall potency rank to each variable, these "Final" ranks are as shown in Table 2. On the basis of these final ranks, Index #1 again identified Variables 9 and 10 as least potent. However, one of these variables (#10) was assigned a rank of 4 by Index #2 and a rank of 5 by Index #3.

Five-Group Case (k = 5). The results obtained in this case very closely parallel those obtained when the number of criterion groups was three. Values of Kendall's W were computed, as well as the chi-square values used in testing the significance of each. The results are reported in Table 3. The average potency ranks of each variable as assigned by each index over 100 runs of the experiment are given in Table 4. Again, Index #1 assigned the lowest ranks to Variables 9 and 10. Index #2 and Index #3 performed somewhat better than in the three-group case but they still failed to consistently identify Variable 9 as one of the two variables of lowest potency. The value of Kendall's W for the (average) ranks as assigned by Index #1 over the four sample sizes was .85. The corresponding values for Index #2 and Index #3 were .95 and .97, respectively. Overall or "Final" ranks were established and are also given in Table 4. Again, Index #1 identified Variables 9 and 10 as least potent. Index #2 and #3 identified Variable 10 as least potent but assigned final potency ranks of 7.5 and 8 to Variable 9.

Insert Tables 3 & 4

Conclusions

The conclusions of this study are limited to a situation in which 1) the k populations are p-variate normal, 2) the k population covariance matrices are identical, and 3) the number of "subjects" drawn from each population is the same. In this situation the findings support the following conclusions:

1) Index #2 and Index #3 can be expected to have comparable reliability and to be more reliable than Index #1 as indicators of the true potency ranking of the predictor variables.

2) Index #1 can be expected to be the most valid in identifying those variables that contribute minimally to the discrimination involved.

3) Given a single run of the experiment, none of the indices can be expected to be sufficiently reliable to be of great practical value in identifying potent variables unless the total sample size is very large. The use of correlations as indicators of variable potency does not appear to be a solution to the so-called "bouncing beta" problem.

FOOTNOTES

[1]The reasons for this restriction are: 1) the first function usually accounts for a major portion of the discriminatory power of the set of predictors, and 2) for each replication of the experiment the number of "significant" functions may not be the same, but there will always be at least one.

[2]Kendall (1955, p. 114) shows that this procedure gives a "best" estimate in a least squares sense.

[3]Tables for the first two indices are presented elsewhere (Huberty, 1969).

REFERENCES

Bargmann, R.E. Interpretation and use of a generalized discriminant function. In R. C. Bose, et. al. (Eds.) Essays in probability and statistics. Chapel Hill: University of North Carolina Press, 1970.

Bock, R.C., and Haggard, E.A. The use of multivariate analysis of variance in behavioral research. In D. Whitla (Ed.) Handbook of measurement and assessment in behavioral sciences. Reading, Mass.: Addison-Wesley, 1968.

Cooley, W.W. and Lohnes, P. R. Multivariate procedures for the behavioral sciences. New York: John Wiley and Sons, Inc., 1962.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, Inc., 1950.

Huberty, C. J. An empirical comparison of selected classification rules in multiple group discriminant analysis. Unpublished doctoral dissertation, University of Iowa, 1969.

Kendall, M.G. Rank correlation methods. New York: Hafner Publishing Company, 1955.

Porebski, O. R. Discriminatory and canonical analysis of technical college data. The British Journal of Mathematical and Statistical Psychology, 1966, 19, 215-236.

Rao, C. R. Advanced statistical methods in biometric research. New York: John Wiley and Sons, Inc., 1952.

Table 1

Coefficients of Concordance, W, and

Associated Chi-Square Values for k = 3

	W	x^2 (9)
Index #1		
N = 90	.113	112.050
N = 150	.177	159.007
N = 300	.302	272.007
N = 450	.381	343.043
Index #2		
N = 90	.182	163.887
N = 150	.259	233.267
N = 300	.288	258.923
N = 450	.418	376.139
Index #3		
N = 90	.189	169.979
N = 150	.265	238.097
N = 300	.299	268.898
N = 450	.425	382.894

$x^2_{.01}$ (9) = 21.666

Table 2

Estimates of the True Rankings

of the Predictor Variables for k = 3

	Variable									
	1	2	3	4	5	6	7	8	9	10
Index #1										
N = 90	8	9	2	4	6	5	3	1	10	7
N = 150	6	7	2	5	8	4	3	1	10	9
N = 300	8	7	3	4	6	5	2	1	10	9
N = 450	8	7	2	4	6	5	3	1	10	9
Final	7.5	7.5	2	4	6	5	3	1	10	9
Index #2										
N = 90	10	2	7	3	8	6	4	1	9	5
N = 150	10	2	4	7	8	6	3	1	9	5
N = 300	10	2	5	4	8	7	3	1	9	6
N = 450	10	2	5	7	6	8	3	1	9	4
Final	10	2	5.5	5.5	8	7	3	1	9	4
Index #3										
N = 90	10	2	7	3	8	6	4	1	9	5
N = 150	10	2	4	7	8	6	3	1	9	5
N = 300	10	2	5	4	8	6	3	1	9	7
N = 450	10	2	5	7	6	8	3	1	9	4
Final	10	2	5	5	8	7	3	1	9	5

Table 3

Coefficients of Concordance, W, and

Associated Chi-Square Values for k = 5

	W	x^2 (9)
Index #1		
N = 90	.069	62.433
N = 150	.143	128.998
N = 300	.236	212.140
N = 450	.351	315.506
Index #2		
N = 90	.099	88.798
N = 150	.121	108.960
N = 300	.191	172.189
N = 450	.318	286.259
Index #3		
N = 90	.112	100.951
N = 150	.129	116.149
N = 300	.208	187.446
N = 450	.332	298.848

$x^2_{.01}$ (9) = 21.666